Mutation analysis pipeline V2: an improved version of bioinformatics pipeline for plant and microbial sequencing analysis

H. Ichida*1 and T. Abe*1

The rapid and coutinuous advancement in the massively parallel, "next generation" sequencing technology enabled a comprehensive analysis of genomic mutations in base-pair resolution in model species, such as rice and Arabidopsis. A commonly used state-of-the-art sequencing platform (HiSeq X, Illumina) produces roughly 2 terrabases of usable sequencing reads per run in less than 3 days. To process, analyze, and extract usable information from such large datasets, a high-throughput bioinformatics pipeline that is specialized for mutation analysis and extensively utilizes parallelization is required. Owing to such dataset volumes, it is not quite realistic to experimentally verify all mutation candidates in each dataset, although the mutation candidates must be accurate for such analysis. We have developed one such pipelines on the "HOKUSAI" parallel computing system, operated by the Advanced Center for Computing and Communication, RIKEN in 2015;[1] further, we have operated and been adding continuous improvements since then. Recently, there has been a great demand to support various non-model organisms in the pipeline. Here, we present a brief description of the major updated version of our pipeline, namely, the mutation analysis pipeline V2 (MAP-V2).

The basic workflow of MAP-V2 is essentially the same as that of the previous version,[2] but the pipeline's internal structure and codes have been redesigned and re-written almost from scratch. MAP-V2 runs quality checking of the input sequencing reads (by FastQC), mapping of the input sequencing reads to the reference genome, and variant calling and filtration to identify mutations induced in the genomes and filter out false positives, which are mainly derived from intra-cultivar variations. In MAP-V2, raw sequencing reads (supported for outputs from both short- and long-read sequencers) are mapped to the reference genome sequences using one of the mapping programs supported by default (BWA, minimap2, ngmlr, and BBMap), sorted, and realigned, and the data are stored in the standard BAM format. Variant callings are performed with a maximum of twelve different programs (BcfTools, BedTools, BreakDancer, CNVNator, Delly, FreeBayes, GATK V4, GATK V3, Manta, Pindel, Sniffles, and Strelka) and merged into a single file. The sequence quality check, mapping, and variant calling statistics are summarized and compiled in an HTML-format report for review. MAP-V2 utilizes a modular structure in each compartment of the pipeline: new programs, workflows, and therefore new organisms can be easily added and integrated later. It supports more programs than the previous version, but this modular structure makes it easy to further extend the pipeline to fit specific requirements for each organism and project. MAP-V2 is currently implemented and run on the HOKUSAI system, and it is capable of transfer onto private and public cloud infractructure with minor modifications, when necessary.

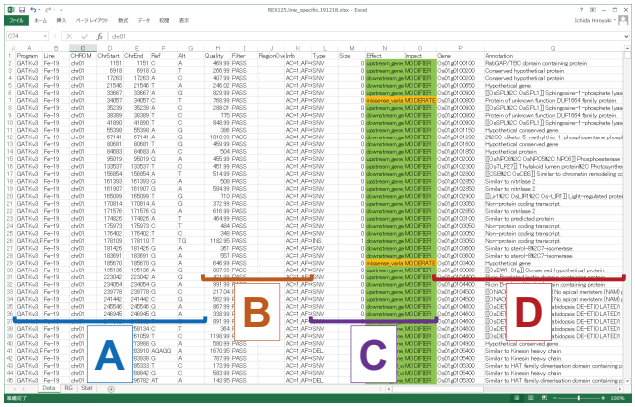

Fig. 1. An example of a MAP-V2 output. (A) Information about the mutation, (B) evaluation, (C) the type and size of the mutation, and (D) gene ID and annotation affected by the mutation.

The most significant change in MAP-V2 for users is the output format of the results. An example of a MAP-V2 output is shown in Fig. 1. MAP-V2 outputs native (binary) Microsoft® Excel® files instead of the tab-delimited text files in the previous version. This change enabled the simulataneous output of multiple datasets in separate worksheets. A result file contains three worksheets that include all post-filtered variants identified from the mutants analysis, mutations that possibly cause the inactivation of gene functions based on prediction from gene structures and sequence changes (candidates of responsible mutations in mutants), and statistics that summarizes identified mutations categorize in mutation types. A variety of formatting options that visually categorize the possible effects of the identified mutations are used in MAP-V2 results.

References
1) H. Ichida *et al.*, RIKEN Accel. Prog. Rep. **49**, 254 (2016).
2) H. Ichida *et al.*, Plant J. **98**, 301–314 (2019).

*1 RIKEN Nishina Center